DOCUMENT RESUME

ED 161 926 TM 007 908

TITLE	Speaking Out: The Use of Tests in the Policy Arena. Proceedings of the New Jersey Conference.
INSTITUTION	Educational Testing Service, Washington, D.C. Education Policy Research Inst.
PUB DATE	77
NOTE	56p.; Papers presented at the New Jersey Conference of the Education Policy Research Institute (February, 1976)
AVAILABLE FROM	Education Policy Research Institute, 1527 New Hampshire Avenue, NW, Washington, D.C. 20016 (no charge)
EDRS PRICE	MF-$0.83 Plus Postage. HC Not Available from EDRS.
DESCRIPTORS	*Conference Reports; *Educational Assessment; *Educational Policy; *Educational Testing; Educational Trends; *Evaluation Criteria; Evaluation Methods; Policy Formation; Program Evaluation; Testing Problems; *Test Interpretation
IDENTIFIERS	New Jersey

ABSTRACT

In February, 1976, the Education Policy Research Institute sponsored this conference on the use of tests in educational policy formation; these six papers were presented at that conference. In the first paper, Jenne K. Britell discusses four points that have historically characterized debates about testing; psychometric and technical issues; issues concerning test use; issues involving the social, economic, and political implications of test use; and moral issues. The second paper, by Joan C. Baratz, focuses on the need to understand cultural pluralism and its significance for the use and interpretation of tests. Samuel Messick's paper discusses tests from an ideological viewpoint in which the appropriate use of a test is considered in terms of its social consequences. The increasing use and misuse of test score data for policymaking decisions at the state level is discussed in the fourth paper, by Jerome T. Murphy. Several innovative methods for establishing educational accountability are described by Senator Jack Gordon, with an emphasis on the need for increased local control in the educational decision-making process. Finally, Samuel Ball discusses the technical considerations involved in the evaluation of public educational programs. A chart listing purposes and methods of program evaluation is included. (JAC)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

SPEAKING OUT

ED161926

THE USE OF TESTS IN THE POLICY ARENA

"PERMISSION TO REPRODUCE THIS MATERIAL IN MICROFICHE ONLY HAS BEEN GRANTED BY
Joan Baratz
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC) AND USERS OF THE ERIC SYSTEM."

U S DEPARTMENT OF HEALTH, EDUCATION & WELFARE
NATIONAL INSTITUTE OF EDUCATION

THIS DOCUMENT HAS BEEN REPRODUCED EXACTLY AS RECEIVED FROM THE PERSON OR ORGANIZATION ORIGINATING IT. POINTS OF VIEW OR OPINIONS STATED DO NOT NECESSARILY REPRESENT OFFICIAL NATIONAL INSTITUTE OF EDUCATION POSITION OR POLICY.

TM007 908

EDUCATION POLICY RESEARCH INSTITUTE

Educational Testing Service is an Equal Opportunity Employer.

COPYRIGHT © 1977 BY EDUCATIONAL TESTING SERVICE.
ALL RIGHTS RESERVED.

TABLE OF CONTENTS

Page

CONFERENCE OVERVIEW

Joel S. Berke v

ISSUES IN TESTING:
AN HISTORICAL PERSPECTIVE

Jenne K. Britell 1

CULTURAL BIAS AND THE USE OF TESTS:
A DOUBLE-EDGED SWORD

Joan C. Baratz 9

VALUES AND PURPOSES IN THE USES OF TESTS

Samuel Messick 20

THE ROLE OF TESTS IN POLICYMAKING

Jerome T. Murphy 27

TESTING AND ACCOUNTABILITY:
A POLITICIAN'S PERSPECTIVE

Senator Jack Gordon 34

EVALUATION IN PUBLIC EDUCATION

Samuel Ball 40

CONFERENCE OVERVIEW

JOEL S. BERKE

Prompted by recent legislative activity and court decisions, as well as by increasing public concern over the declining quality of the public school system, New Jersey has been confronted with the task of redefining its educational priorities and policies. In the course of public debate a series of controversial questions have been raised concerning the use of tests in education and in policymaking on a statewide level: For what purposes should tests be used? What part should they play in the allocation of financial resources? Should minimum standards in assessment programs be used? If so, upon what criteria should they be based? Who should set them?

In February 1976, the Education Policy Research Institute, a research division of the Educational Testing Service, sponsored the New Jersey Conference in order to provide an informal setting in which to discuss these types of policy questions. Designed as a forum for discussion and information, the conference involved participants with a broad range of perspectives and areas of expertise concerning education and public policy. Included were representatives from the State Department of Education, the Governor's Office, the legislature and the legislative staff, the New Jersey Education Association, the School Board Association, the National Conference of State Legislators, minority group associations, businesses and other interest groups. The varied nature of this group provided a unique opportunity to examine the policy issues involved in improving the quality and equality of education in New Jersey, especially as they relate to the use of statewide testing.

The breadth of the issues examined at the conference is apparent from the papers contained in this publication. Perhaps most significant is the fact that despite the differences in the problems addressed by each speaker, there is an underlying commonality of urgency, tempered by creative alternatives for the future. Taken individually or together, the result is an interesting and thought-provoking way of viewing the problems common to any group or individual concerned with increasing the efficiency of education in the United States today.

The first paper, delivered by Ms. Jenne Britell, discusses the issues of testing from an historical perspective. It is, perhaps, disheartening to learn that the fundamental issues in the debates over testing have

changed little in the 70 years that mental tests have been used in education, and further, that the questions that plagued the early proponents of testing remain virtually unanswered today. Ms. Britell cites four major focal points that have characterized the history of debates about testing: psychometric and technical issues; issues concerning test use; issues involving the social, economic and political implications of test use; and moral issues. There have been variations, of course, as to which of these issues took precedence in discussions, as well as in the perspectives in which the problems were discussed, and the group of people who led the debates, but the consistency of the issues themselves is so remarkable that it is cause to wonder if any progress has been made. Indeed, the subject matter and focus of the presentations of this conference only serve to reinforce this consistency by clearly echoing these major issues and concerns. There is, however, a significant difference in the manner in which these problems are approached. Evident is a new type of frankness about the limitations of test data as a major lever for producing educational change.

A major area of concern in the current controversies over tests, is the cultural bias issue. In an incisive commentary on this problem, Dr. Joan C. Baratz presents the argument that the concerns and debates over the large failure rates of minority children on standardized tests have been misdirected. All too often, the issue has been discussed and acted upon in ways which define the problem in terms of blame and responsibility: The tests are inaccurate; the teachers are failing in their jobs; the parents are responsible; the school system is inefficient; society misuses the tests; minorities are inferior. But Dr. Baratz argues that culpability is not the fundamental issue. What is more important is the recognition of cultural differences and their role in the educational system. Illustrating the cultural difference model with examples from Black language and culture, Dr. Baratz demonstrates that the diversity of cultures are manifest in many subtle ways. She argues that the educational system can benefit only through an understanding of the meaning of cultural pluralism and its significance in regard to the use and interpretation of tests.

In addition, Dr. Baratz points out that it is not a question so much of whether tests are culturally laden or not, but rather whether this fact is necessarily incongruent with the goals of education. It is not the use but the interpretation of tests that is the difficulty. Baratz demonstrates that those who feel that tests should be discarded because they are culturally biased, are perhaps as naive as those who do not acknowledge that cultural differences exist.

Dr. Samuel Messick's presentation focuses upon tests from an ideological viewpoint. He emphasizes that there is a fundamental distinction between the adequacy of tests and their purposes which must be

realized in debates over testing. The question of adequacy involves the evaluation of a test in terms of its psychometric accuracy, and more important, in terms of its meaning. This is a scientific and technical question which social scientists have a well-established means of assessing. In contrast, the second question concerning the purposes of tests, involves a more subjective and philosophical approach. The appropriate use of a test must be considered in terms of its social consequences as they relate to the more ideological concerns of societal values. It is Dr. Messick's belief that until some of these latter issues are resolved, the use of tests to support the educational system will most likely make little progress.

One area of controversy which has received considerable attention in the past decade has been the increasing use of test score data for policymaking decisions at the state level. Intended to improve the efficiency and effectiveness of the education system, assessment and accountability programs often suffer from technical and conceptual problems. Dr. Jerome Murphy, in discussing the relationship of tests to policymaking, cites examples drawn from various statewide assessment programs that indicate that their usefulness has been somewhat negligible.

The voluminous quantities of data generated by statewide standardized tests all too often indicate only quantitative differences between schools and school districts. For the policymaker interested in knowing *why* a certain school scores consistently lower, *what* the underlying reasons for the test score disparities are, or *how* to improve those schools identified as having a problem, this type of data is clearly inadequate. It is Dr. Murphy's contention that more reasonable expectations need to be established for the use of tests in policymaking. Further, he believes that there are alternative methods of evaluation which do not involve tests that might prove more helpful in making decisions about the educational system.

Related to the issues of evaluation in education is the concept of accountability. Senator Jack Gordon discusses this problem from the viewpoint of legislators and citizens who are concerned with improving the educational system. While accountability is most often measured in terms of test score data, Senator Gordon emphasizes that there are alternative methods of validation which ought to be considered. Drawing upon his experiences in the Florida state legislature, he describes several innovative methods of accountability which have been successfully implemented in his state in recent years. Perhaps most significant, these innovations emphasize the need for increased local control in the educational decision-making process. In essence, they herald a new philosophical approach to the problems of accountability by allowing increased flexibility for the school and providing

the opportunity for community members to become directly involved. Gordon indicates that to effect change in the current system of education, the artificial elitism that has characterized the profession must be discarded in favor of an approach which maximizes the resources of the community and which recognizes the viability of suggestions made by outsiders to the educational process.

The final paper, by Dr. Samuel Ball, discusses the technical considerations involved in the evaluation of public educational programs. He emphasizes that there is a fundamental distinction between the evaluation of individual students and the evaluation of programs—an important factor in determining the methodological approach and purpose of the evaluation process. Arguing that the essential reason for evaluation is decision-making, Dr. Ball urges that programs must be developed in terms of the purpose which they are to serve for decision-makers. In his view, there are five such purposes: to contribute to decisions about program installation; to contribute to decisions about program continuation, expansion and/or accreditation; to contribute to decisions about program modification; to obtain evidence in support of or in opposition to a program; and, to contribute to the understanding of basic processes. Dr. Ball presents the considerations involved in each of these purposes. Most significant for this conference is the relatively minor role Ball assigns testing in the evaluation process. Indeed, it is Dr. Ball's belief that although tests may be a viable method of assessing the quality of educational programs in certain instances, more often there are other methods which are more appropriate and which should be considered seriously.

There has been a growing, if not unfortunate, trend to identify objective data as a base for decision-making in education. This has led to increased emphasis on the use of test data in the education policy process. While the use of test data in education is a critical issue today, it is evident from these papers that there are severe limitations politically, technically, and ethically to the exclusive use of such information. It is our hope that this publication will be useful to decision-makers who must consider the role of test data in policy formulation. In addition, it is hoped that this publication will be valuable to those who are seeking constructive alternatives and ideas which can be incorporated into positive methods for the improvement of our current educational system.

Joel S. Berke
Education Policy Research Institute
Director

ISSUES IN TESTING: AN HISTORICAL PERSPECTIVE

JENNE K. BRITELL

The use of mental tests has played an important but controversial role in the recent history of education in the United States. I would like to begin my discussion with some general descriptions of the major trends and issues that have characterized this period. Later I will provide some specific examples that illustrate these points, and then finally offer some conclusions that can be derived from this historical perspective.

For approximately 70 years, standardized tests of achievement and aptitude have been used in education. Throughout that period, the literature reveals substantial confusion about what tests are intended to measure, what indeed they measure, whether they measure it well, and whether the quality measured relates to anything important. Similarly, confusion reigns over the definition and use of terms such as ability, aptitude, capacity, achievement, intelligence, mental age and I.Q., despite the continued efforts of professionals to define them.

There have been multiple uses of tests in education throughout their history. They have been used for instructional purposes such as ability grouping, causing a constant debate through the decades as to whether homogeneous or heterogeneous grouping is more desirable. Tests have been utilized for grading and more recently in the credit by examination movement. In addition, they have played an important role in guidance, college admissions, industrial placement, professional certification and military manpower classification. As a result of these various uses, tests have been viewed and treated not only as a pedagogical or vocational tool, but as a sorting instrument as well. This latter role has been severely criticized and has played a major part in the controversy surrounding tests.

Historically, there are two trends which have consistently characterized the years of test use. First, tests have been of particular importance during periods of extraordinary educational, economic, political, or social change, such as war or population booms. For example, when there have been more postsecondary applicants than places, tests have become an important part of the admissions process. Although we tend to think of this as a characteristic of the fifties and sixties, it marked the twenties, after the first World War, as well. Similarly, the expansion of elementary school populations has increased the

use of tests to improve teaching effectiveness and/or to increase administrative efficiency through ability grouping. Again, in the twenties, with the sudden influx of a large number of non-English speaking immigrants entering the school system, New York City administrators turned to testing to improve the process of education.

A change in national priorities can also increase test use. In the fifties, the post-Sputnik era called for a talent search, similar to the push in the twenties to find superior children, because they were considered a national resource which must be conserved.

Test use has been fostered as well by interest in the whole child. Several times in the history of education in the U.S., testing has been used to individualize instruction, and individualization has been presented as a reason to continue testing. There is considerable irony here since much of the criticism of testing centers on its perceived tendency to diminish concern with the whole child.

Inevitably the growth and importance of testing in such periods produces growth in concern about its use—tests are suddenly everywhere and in everything. Concern about the effects of over-reliance, overemphasis, and the consequently diminished importance of, or attention to, other issues begins to dominate discussions. Thus, a cycle of increasing interest, increasing use, increasing concern, increasing scrutiny, and for a time, diminished use, tends to exist.

Second, the issues enveloping testing have been consistent and, unfortunately, in some instances, remarkably resistant to resolution. Testing uses appear to cluster in four major areas:

- *Psychometric or technical issues:* These include the validity of a particular measure, the definition of what it measures, such as intelligence, aptitude, or reading, and the relationship of a test to the criterion, such as grades.

- *Test use:* These issues include the usefulness and limitations of a measure for educational, vocational, or administrative purposes.

- *Social, economic and political implications of test use:* These focus on certain policies which make use of test-produced information. What is the effect of putting all the "superior" children in one class or one track? What is the effect for the school's efficiency? What is the effect for the taxpayer's dollar? What is the effect for the child's development?

- *The moral issues:* These issues bear on basic values that the society holds and the ways in which the test information supports or contradicts them. Some examples are: equal opportunity and the use of tests for selection; the capacity of individuals and the potential for

change through education; the creation of an elite, based in part on intellectual skills (as evidenced by test scores), and a democracy of equals.

Nineteen-seventeen marked the first large-scale use of mental tests of ability in this country in contrast to the earlier achievement tests in specific subjects. Even before that time there had been a concern with, and a recognition of, all these issues. This point should be stressed, because there has been a tendency in recent years to emphasize the lack of concern for some of these issues by those who develop or use tests. The literature reveals, however, a constant effort at refining testing instruments, and specifying what they should be used for. Central to this psychometric issue has been the question of whether a test measures what it is supposed to measure. There has been a constant attempt to limit, to criticize, and to look at tests in connection with other techniques which might be better. For example, the concern over these issues was evident at a symposium sponsored by the *Journal of Education Research* in 1935. The questions were asked: Are tests determining what should be taught in schools? Are teachers teaching to tests? Is this good (in setting a standard)? The consensus was, as it would be today, that teaching should not be to the test. While measurement was acknowledged to have a place, it was considered by no means the only important factor. Further, there has been a continuing concern and recognition that the type of information produced by tests could lead to questions on the basic values in the U.S.

William Bagley, had earlier recognized this issue of basic values in testing when he wrote in 1922: "Democratic education is fundamentally not interested in the possibility that this individual may surpass that individual, but rather in the absolute advancement of all."

The use of tests to separate children into one group or another has always raised questions about a basic democratic principle, at the same time that research has indicated testing's utility in teaching or administration. Lewis Terman, a believer in certain limitations of training and a leading psychometrician, stated in 1928 that, "taking issue with a democratic principle about what education could accomplish was like taking issue with the right of the people to control in a democracy." Subsequent speakers on ability grouping acknowledged the same issue.

Participants in these discussions throughout the last 70 years have included individuals responsible for the development of measures, those who taught the use of such measures, those who used them—principals, superintendents of schools, professors of educational administration—and the public, parents, legislators, and taxpayers. Although we tend to think of questions of efficiency and effectiveness as recent issues of the 1960s and 1970s, taxpayers in the 1920s were

equally concerned that they were not getting their money's worth from the schools. Teachers and principals were using test scores to show them that, indeed, they were. John Dewey was very concerned about such usage because he feared that the test and test score became ends in themselves. He felt that the teachers saw test scores as indicative of limits of ability, not guideposts for future effort by the educator.

Although the issues have remained constant, various groups have raised these concerns at different times. There has been a tendency on the part of those who develop and use tests to deal primarily with the first two categories, the psychometric issues and usefulness issues, although individuals in this group have also been concerned with the second two categories. Frequently the social critics have emphasized the social, economic, and political implications of testing and the broader issues of democratic values, with sometimes little understanding or recognition of the merit of the first two categories. Again, these are generalizations and are not uniformly true. There is an unfortunate paradox of criticism from the perspective of democratic values, for although testing is viewed as harmful, the original intent of the testing movement in education was to improve the educational process, to make the system more equitable, and to heighten efficiency.

History supports these generalizations. In 1917, Robert Yerkes, then President of the American Psychological Association (APA), sent a telegram to Washington offering the knowledge of the APA to assist in the recruitment efforts of the U.S. military. Within three months, Yerkes and almost everyone who would have been in a "Who's Who of Psychometric Measurement," were busily at work classifying 1.7 million recruits. "Everyone" in this instance, included Lewis Terman, Arthur Otis, Edmond Boring, Yerkes, L. Thurstone, Rudolph Pinter, E. L. Thorndike and others whose names frequently appear in testing literature. By the time they had finished classifying the 1.7 million, the war was over, thus precluding a large-scale extensive demonstration of the tests' validity.

The results of this testing effort (in which the Alpha was used for literate English-speaking recruits, and the Beta for non-literate and non-English speaking recruits) were published in 1921 by the National Academy of Science. This was the first large-scale publication of a major testing effort in the United States. It led to the first large-scale misinterpretation of test results as well (including the theory that the mental age of Americans generally was 14 and that Philadelphia had an average mental age of seven). *The New York Times* pointed out that Philadelphia, cradle of liberty and source of so many cultural, literary and national efforts toward improvement, had better consider what was happening to it and perhaps, offered the *Times*, what was needed were some good psychologists who could advise Philadelphia on how to deal with its problems!

Just as there were those who said they knew it all the time, there were those, like Walter Lippmann, who took issue with the alleged testing results, and concluded, there was something wrong with the theorists and the measurement rather than with the population of the United States. Although Lippmann's and Dewey's concern about the erroneous concept of mental age and the misinterpretation was shared by psychologists such as Frank Freeman, psychologists primarily devoted their efforts to psychometric issues such as the improvement of the technical properties of tests, and to an emphasis on more accurate use, assuming greater understanding would eliminate misuse and misinterpretation.

The mental age furor, if not ignored, became submerged in other equally important concerns of education during the 1920s. The New York City public schools, for example, were confronted with large groups of immigrant children who did not speak English. For that reason, in 1922, New York turned to an intelligence test and, on a trial basis, tested 40,000 students in an effort to reduce the waste occurring as a result of mismanagement of education. Subsequent to that successful trial basis in which it was found that children learned better when classes were separated according to ability, intelligence testing was instituted throughout the city, a practice which other cities followed shortly thereafter.

At the same time that this heterogeneous growth occurred in urban elementary school populations, a growing number of people, primarily males, desired a college education, since one earned more money and achieved greater stature with a college degree. Suddenly, colleges received more applicants than they could handle; there were not enough buildings; the professors refused to teach; and an ensuing revolt among educators occurred at the college level. Columbia was the first to move to the use of a psychological examination as a part of the admissions process. They also continued to use the essay examinations developed by the College Board several years earlier. The point was carefully reiterated that this examination would be used *in addition* to the essay examinations, the grade-point-average, recommendations, and character references.

The New York Times observed that Columbia's action represented a change in a basic value: society had always assumed colleges were to be of service to the *community*. If they were to be of service, the *Times* observed, it was valid to ask whether they were of more service if they provided *some* education for the *many* rather than a *great deal* of education for the *few*—even if the few were more able in terms of the criteria that Columbia was using, and if educators preferred to teach only able students. The *Times* concluded, "In a sense, we are losing a cherished American ideal." By the end of the twenties, however, many Eastern colleges—ivy league institutions, like Colgate (whose presi-

dent had already decided that there was not much point in trying to reach most of the 18 year olds), and the City University of New York —had moved to the use of psychological examinations. (Initially, a version of the Army Alpha was used until other aptitude examinations replaced it.) Although the notion of testing to identify the elite for colleges was firmly established it was not uniformly accepted or practiced. Many institutions there raised a cry still heard today: What about the obligation to all?

The movement to use these examinations was matched by continued efforts to improve them. Substantial research was conducted by those involved at the university level, by individuals, such as E. L. Thorndike and Terman, and by school systems and their newly-formed bureaus of educational research. Philadelphia, New York City, and most state departments of education had formed bureaus which studied test use in areas such as ability grouping. The issues discussed in this period of the 1920s were primarily psychometric and educational. In the thirties, the orientation became a renewed emphasis on the categories of social and political implications, and the broader concerns of a democracy. This was in part because of the progressive education movement and its emphasis on the whole child, in part because research was moving forward, and in part because people began to explore the consequences of too much reliance on any pedagogical tool. However, in the 1936 *Yearbook of the National Society for the Study of Education*, ability groupings was still favored as the way to encourage children and to improve their work.

The thirties were also marked by recognition and corresponding research on culture and environment. Although we tend to think of the concerns of special interest groups and of issues such as cultural bias as relatively new, the thirties were characterized by substantial research in areas such as the effect of bilingualism on the ability to learn, on intelligence and reading test scores, and on black-white differences. In an attempt to explain the findings of the research, both environmental and heredity-related explanations were advanced. There was a study done in Trenton, New Jersey which seemed to show that the performance of Italian children on a test of reading achievement at the third grade level was the same regardless of whether English or Italian was spoken at home. The New York City Italians took great issue with the use of tests in the classifying of their children and the use of such tests with Mexican-American populations in other parts of the country provoked a similar outcry. The *Journal of Negro History* carried many articles on test performance as well.

The 1930s were also characterized by an increase in the nature-nurture controversy: whether heredity or environment influenced behavior most. Ten years earlier the evidence had seemed to fall on the side of nature, but by 1938, nurture, the effect of the environment, seemed

to be in the ascendency. This stance was supported by the work of researchers, such as Beth Wellman at Iowa who had found that with an intensive nursery school program for orphans, age three to five, I.Q. scores could be raised from 10 to 30 points on the standard I.Q. tests given to children in the lower grades. Children who were not given this opportunity seemed to lose 10 to 30 points on a test of performance. Although the methodology and research design of the Iowa group were strongly criticized, this experiment may well be the first instance on record of organized preschool intervention and its effects.

In addition, there was questioning by individuals like the psychologist Frank Freeman. Freeman's comment, that the use of tests fostered "the comparison of children one against the other," is a remarkably contemporary comment. In 1937, the Carnegie Corporation sponsored a study on "the real danger of examinations," in which it observed that society needed to face the question of whether it wanted to produce everyone in one mold, whether it could, and if not, what alternatives would be acceptable.

The concern with I.Q. and particularly questions of its constancy and susceptibility to change, endures from then until now. Clearly the resolution has implications for the kind of educational program one supports—as we have seen in the response to the arguments put forth by people like Arthur Jensen and Christopher Jencks.

In the 1940s, there was a lessened emphasis on discussion of tests, probably because of their accepted use in World War II classifications and because of a smaller population on the college level. However, the fifties and sixties saw renewed interest in the discussion because an increasing population again entered elementary and secondary schools, as well as sought admission to postsecondary institutions. An emphasis on talent also made tests more important because of their utility in indicating who might succeed. Ability grouping in the lower grades appeared again and questions of accountability and efficiency reappeared as well.

In the 1960s, the Russell Sage Foundation undertook a study similar to the earlier Carnegie Corporation study to see what the attitudes were on testing of ability. Although they attempted to separate out intelligence, aptitude, and achievement tests in their descriptions, the confusion of terms still existed. The Sage study indicated several things: people felt that testing was useful; the abuses of the tests were of concern; and criticism was diverse but not clear-cut or consistent.

The last years of the sixties and the beginning of the seventies have been marked by an increasing concern with the social, economic, and political implications of using tests to place a child in a certain program, and with the effect on equality of opportunity of the use of measurement within the educational process. Today this concern characterizes contemporary discussions within the psychometric pro-

fession, as well as without. Yet the similarity of topic and acuity of concern are frequently lost on the varied participants and not evident in the rhetoric employed by all.

Everyone is concerned: if one analyzes arguments of both advocates and critics of testing over time, their concerns are not that different, and there are not polarities when the discussion initially appears: both agree that the educational process needs improvement. While tests usually are brought in to eliminate one set of problems, they sometimes appear to create another. It is truly unfortunate that the commonality of interest and concern—for children and for the process—is not apparent to the parties to debates on the use of tests.

One becomes aware that there is no consistently high level of knowledge about tests possessed by everyone affected by the process of testing. What tends to happen, then, is that one never gets a dialogue among people whose interests may be similar—improving education for children; making sure finite resources are used most effectively; trying to ensure that children are having a positive educational experience; attempting to make reasonable equitable administrative decisions.

Perhaps most necessary—and we may be moving toward it through meetings such as this—is that all parties to the debate make a stronger, more responsible effort to share information about the uses and limitations of tests, to respect the legitimacy of one another's interests, and to understand various views. Such efforts could produce more appropriate tests, more accurate and effective use, and better scrutiny by the public.

CULTURAL BIAS AND THE USE OF TESTS: A DOUBLE-EDGED SWORD

JOAN C. BARATZ

One of the major concerns of this conference is the large rate of failure of many minority group children in the current educational system. We know this: We do not need tests to ascertain this. We know there are many dropouts, and students who remain in school are very dissatisfied with the manner in which they are being educated or "miseducated."

We also know, however, that tests dramatically and continually demonstrate that failure. I believe it is not the tests, but the anxiety about the test results that is the controversial issue. That anxiety concerns both how the test information is interpreted and what the information is used for.

There is a great deal of legitimacy surrounding the fears of people about the use to which test results are put. This is not merely an academic debate among "objective social scientists." If the past is an example, the information generated from tests has been used against the individuals who took the tests. Ms. Brittel gave us the example of the furor created over test use in the twenties when, on the basis of I.Q. test results, the populace of Philadelphia was said to have a mental age of seven years. While that might sound absurd and amusing today, and it might appear that such information would have been ignored, if not ridiculed, by policymakers in the twenties, such was not the case. The information generated from those tests was used, and used against the individuals from whom the results were gathered, in the sense that restrictive immigration laws were passed, and friends and families of immigrants to the U.S. were prohibited from coming to America.

Tests are used to exclude, classify, and label people. It comes as no surprise, therefore, that those who are excluded, pejoratively labelled or misclassified would find objection to tests. This is especially the case, when they are excluded from the decision-making process regarding the consequences of such labelling. Experience has demonstrated to many individuals, most especially minority group members, that despite the claims of the test givers that tests will help educators by providing the information they need to more efficiently and effectively educate children, tests have been harmful to them. Tests have kept them out of jobs, tests have tracked them into dead-end

educational programs, tests have subjected them to ridicule—they have been used to blame students for their failure—"you are too stupid to learn."

The use of tests to assign blame is cause for great concern. The fact that tests demonstrate great failure rates provides an occasion for many who are interested in efficiency and effectiveness, and even those who have a more humanistic approach to education, to look around and say: "Who's to blame for this? Who's responsible for this? We've got to do something about it."

There have been two major groups in the past few years that have been prime targets for blame, for having the information about educational failure used against them. These two major groups have been (1) the child and his family; and (2) the school system, and, in particular, the teacher. When someone is blamed, it is very natural to attempt to explain away that responsibility. These two groups, families and educators, are no exception, and they have developed explanations to deal with the blame that has been put upon them.

Teachers and parents have not always been allies concerning the testing controversy, and, in fact, at one time tests were greatly endorsed by teachers as part of the scientism and elitism of the profession. However, with the movement toward blaming teachers for the failures of children, the teacher organizations began to have anti-test sentiments. Consequently, they have joined many parents in endorsing, as one explanation of the high test failure rates, the cultural bias of tests. This viewpoint usually infers (1) that if the tests were not biased, the differences in performance of minority group children would disappear, and (2) since the tests are biased, they should not be used.

While the cultural bias model of tests may well explain a good deal of the poor performance on standardized tests of children from culturally different homes, it is certainly not a justification for abandoning such tests, especially if one endorses as one goal of public education the preparation of all children to function in the mainstream cultural system. The standardized tests with their content and assumptions reflecting mainstream cultural traditions cannot be viewed as a measure of native aptitude for children who, until they reach the educational system, have had little contact with such cultural norms. However, assuming the goal of the public schools to be that of assisting the child to compete in a pluralistic American society that demands some knowledge of mainstream norms, it behooves educators to view such standardized tests as guideposts to determining how well the school has succeeded in teaching culturally different children about the mainstream. The tests do not, therefore, reflect the "educability" of culturally different children so much as they reflect the extent to which the schools have succeeded in educating culturally different students.

10 Let me explain what I mean by cultural difference and how such dif-

ferences are likely to influence performance on standardized tests. To understand the cultural bias of tests it is first necessary to examine how schools, as preservers and transmitters of the myths and rhetoric of society, have viewed the varied cultures that comprise American society. In the last 10 years we have seen a profound change in the society which has moved from an assimilationistic view of cultural diversity to a pluralistic one. This is perhaps best reflected in the shift in the metaphor of American society as a "melting pot" to the notion of American culture as a "tossed salad." The move from a position of America as the melting pot is to a large degree the result of assertions by both minority group spokesmen and social scientists that there are many cultures in the United States that interact with each other but still maintain their distinctiveness, and that these diverse cultural groups function together in the way that a salad "functions"—the whole is greater than its parts, and the parts are distinct; a tomato is different from lettuce!

Basic to these metaphors is an underlying dichotomy of the notion of "difference." The melting pot idea was a notion of assimilation. Each group of people came to America's shores with a distinct language, distinct group habits and distinct attitudes about institutions, whether they be education, government, or what have you. The melting pot theory posited that by virtue of coming together to this golden land of opportunity and free enterprise, each group immediately merged with other groups and picked up the mainstream culture, which was the product of all that commingling.

Let us examine this thesis more closely. A melting pot does two things. It takes diverse metals, and from these diverse metals extracts the strengths of them all in order to build the most desirable metal possible from the combination. At the same time, it washes out the weaknesses. In relation to society, the metaphor suggests that the result is a unique society, the "American mainstream culture," which has the "strengths" of all the cultures that contributed to it, and, of course, is devoid of all the "weaknesses" of each distinct culture.

Inherent in this concept was the assumption that if a cultural group had been in the United States for awhile, and still had not managed to "wash out the impurities" and differences of their culture, there must be something wrong with them. This made a discussion of difference extraordinarily difficult, because it implied some inherent, immutable, uneducable aspect to individuals that could not be erased by virtue of living in America. Discussion of ethnic differences was viewed as an embarrassing admission of failure to "make it" in the American society. This discussion was further confounded in the thirties, forties, and fifties by the spectre of Nazi Germany's assertions of cultural difference being a sign of genetic inferiority. Thus, discussions of difference became linked to inferences of the failure of the American demo-

cratic system and also became equated with racist assertions. Within the melting pot myth, the notion of equality became one of sameness. Difference was equated with defective, and the real argument was not whether an individual had a legitimate culture or not, but whether the individual behaved the way he did because he was defective inherently, that is genetically inferior, or whether it was because of his environment. Thus, the issue of differences became linked to the well-known nature-nurture controversy.

A liberal/conservative split developed over the issue. The liberals, on the one hand, denied that there were differences and explained that "those people would behave just like 'us' (meaning those of us who had melted into the mainstream culture), it it weren't for the fact that they lived in poor housing or did not eat right, or the like. If they didn't have so much trouble economically trying to survive and the environment was not so oppressive, they would be just like mainstream Americans (implying that then they would be okay)."

On the other hand, the conservatives recognized the differences and used them as justification as to why "those people" could never be like "us." They posited genetic inferiority to explain the different behavior. The interesting thing was that regardless of the viewpoint, liberal or conservative, both groups were saying virtually the same thing. Whether the cause was immutable or environmental, the point was that both liberals and conservatives viewed the behavior of culturally different people as defective and pathological.

The problems with trying to settle a controversy that was based on the same underlying assumption became evident in the sixties. The environmentalists, trying to change the oppressive conditions, found that they could not change them fast enough. Intervention programs that started at the school level did not work. So the war on poverty moved to the preschool level and when that didn't work, it moved to pre-preschool by trying to teach mothers how to be better mothers. Finally, it moved beyond that to trying to teach mothers how to eat properly before motherhood, and, lo and behold, the environmentalists had backed right up into the people who took the genetic positions; a delicate point that embarrassed everybody involved. In fact, the environmentalist model had lost so much credibility that in 1970 it was all right to publish an article like Jensen's "How Much Can We Boost IQ and Scholastic Achievement?," which five years earlier could have been peddled to every journal, no less to the *Harvard Education Review*, and never been printed. In the chaos of trying to understand what was going on and trying to deal with the issue of difference, we had gotten into such a confused state, that the only option open was to recycle and start discussing the genetic/environment issue again. But this time, the metaphor began to change and people began to focus on the tossed salad assumption, which discussed differences from a

cultural perspective rather than in terms of genes or environment.

This brings us once again to the issue of cultural bias in testing. Culture means not only different dressing patterns but essentially a whole system of learned behavior, involving such features as attitudes, belief systems, and language systems. For purposes of discussing cultural bias in testing, I am going to present examples from black culture. I do this for several reasons: (1) I know a little more about that than other cultures; and (2) the bilingual issue is a little bit too "simple" because in all the discussions of bilingualism, going back to the thirties, language of national origin has been recognized. It is easy for us to say, of course, if the child speaks Italian, he has a language. It may not be a language that he needs in public school today, but we recognize that he speaks the language.

With the black culture that was not the issue. There was no specific and immediate flag orientation or nationality orientation, and in contrast, there had been a deliberate denial of black culture. In fact, any time a distinctly black cultural pattern was incorporated into the mainstream culture, we were likely to assume that when blacks manifest that behavior, they learned it from "us." There was no recognition of the importance and the contributions that blacks had made to the mainstream culture, because they were denied the viability of a culture that had things to contribute.

In discussing cultural bias in tests, I will first focus on language, because language is a part of all the tests, is generally the medium of instruction, and is something that is not only personal and obvious but also right out front. In addition, there has been a good deal of work on the structure of black English in the last ten years.

I became interested in this issue a few years ago. With the ink still drying on my degree, I had been told that it would be a good idea to get something published quickly so that I would be employable in the future. At the time I was employed at Howard University in a research project dealing with language development. A young woman at Harvard had done a very clever study to demonstrate how and when white, middle-class youngsters developed the plural and I decided that I would do a similar study with black children. Learning the plural is not as easy as it might appear. It is not always a simple matter of adding an "s." The plural has several different forms as well as several different sounds. For example, when a word ends with a sound that is unvoiced, like a /p/ or a /t/, to make the plural you add an /s/, but for a word that ends with a sound that is voiced, like an /n/ or a /g/, you add a /z/, as in /bunz/ (buns) or /bugz/ (bugs). This is a complicated rule, and the child has to learn the difference between when to use /s/ and when to use /z/, as well as what to do if a word ends in a sibilant sound like /s/, as in "dress." In such instances he has to learn to say /ez/ (dresses).

The test that had been devised by Jean Berko at Harvard consisted of pictures of nonsense figures. The tester would present a picture and say, "Here's one wug, now there are two, there are two _____," and then she would see if the kid said /wugz/. When I administered the test to black children, they often responded, "There are two /wug/." These results matched those of most educational psychologists who said black children were "language delayed," and among other problems had no concept of the plural.

Now, I happened to live in a neighborhood where there were lots of black children, and I knew if I promised them two pieces of candy, I could not give them only one. So I knew they knew something about the plural but I could not understand why it did not show up on the test. The answer came when a linguist, William Stewart, explained to me that if you examined the language that black children were speaking, you would see that the plural did not have to be identified by /s/, /z/ or /ez/, if it was already marked by a number. In other words, "I got five cent" was perfectly good "English," because the five told you that it was already plural. It was not that these children had no concept of the plural, but rather that they had *different* linguistic rules for expressing it.

With this insight into the problem, I went back and administered the tests without giving them the information "here's one and now there are two," and instead just showed the children pictures and said, "This is a wug, now what are these?" I got /wugz/ all of a sudden, and I realized that the basic problem had been one of confusing a manifestation of the plural as it is indicated in standard English, for the concept of the plural itself.

This was the beginning of my studies of black language. We began to document the very orderly, very well developed, very highly structured and very predictable nature of black English grammar, as well as to describe how it differed from standard English. Our studies began to explain things like double negatives, which were perceived by school teachers as an inability to learn, and which actually represented a tremendous misknowledge on the part of teachers about language. Standard English is an exception, most languages have rules employing multiple negation. The point about black language, as with all languages, was not that the features of the language were unique, but that the grouping of the features of the language was what constituted the difference. We discovered that the educational problem was that of match and mismatch, between what the black child was doing and what the perception of people using the tests with these children understood about language ability. *The testers were misreading the black child's inability to perform a standard English test as an inability to generate a language, rather than as an inability to generate standard English—a very significant distinction!*

This kind of mismatch relates to cultural bias in several other ways, and helps to explain why these kinds of tests are problems to certain groups of children. First, is the issue of the attitudes involved in test taking. For example, the Jewish culture is an extraordinarily child-centered one. Children in that culture are able to interrupt parents at any time and "perform." They are not only able to do this, it is expected behavior. For example, when I phone my sister in New York and I want to do some quick business with her, she says, "Wait, talk to Emily." That's nice but Emily happens to be 11 months old! Nonetheless, I "talk" to Emily. My niece is held next to the phone and I repeat, "Hello, Emily, etc." She drools at the other end. But the point is that even before she can talk she has already learned that she is supposed to interact with adults. This type of atmosphere is very different from the black child who does not grow up in a culture where he is socialized early and expected to perform immediately and verbally in front of adults. In fact, regarding black child rearing patterns, it is not an inappropriate designation to repeat the old adage that is looked upon as a part of the American ethic: "Children should be seen and not heard." This is more in keeping with the young black child's style of interaction rules in relation to adults. Therefore, when these two children—black and Jewish—are put into a testing situation, it is not the same kind of situation for them, no matter how "friendly" the examiner, white or black, is.

One child has been primed so that when you show him a little ball or a little block and say, "Tell me all about it" he is ready to tell you *all* about it. The other one has learned the rules too, but his rules of interaction are not display, but are more appropriately expressed by, "I don't know," or, "It's red." That's appropriate, he hasn't bothered you and you haven't bothered him. That response represents learned behavior, not ignorance or a lack of knowledge. The result is a whole different set of attitudes toward test taking.

Another example concerns reading and attitudes toward tests. Many black children come from a culture where reading is not a central concern, and where survival can be very effective without reading. Thus, the notion and meaningfulness of reading is not going to be the same. And, in fact, as an aside, I would say that one of the reasons that a lot of middle-class kids learn to read with the "See Johnny Jump" texts that we give them is because they make such drivel meaningful at another level by saying, "Gee, I'm just like dad, I can read this book in front of me." But to a child who doesn't have that kind of meaning applied to the experience, such "reading" materials are totally empty.

Another type of bias problem concerns the actual content of the test. For example, achievement tests have reading passages where Johnny writes home to his friend in fourth grade about his trip to the Alps, or how his mother made fondue for dinner, or something equally absurd.

Clearly, in terms of vocabulary and types of experiences, this is a blatant form of cultural bias in favor of children from mainstream, middle-class homes.

Another problem occurs because of different meanings given to words. When I was at Howard, I spent some time with a class of preschoolers. In those days it was believed that milk was good for everybody—subsequently, Johns Hopkins has shown that some children get stomachaches from milk due to chemical imbalance. But in those days you were supposed to say, "Be a good boy or girl and drink your milk." Coming from my Jewish background of making a fetish of food and eating, I said to one little boy who hadn't drunk his milk, "Don't *waste* your milk." The child looked bewildered and said, "I ain't wasting it." "Well," I said, "then drink your milk, don't waste it." "I ain't wasting it," he insisted. We went back and forth, getting no where until it occurred to me that to him "waste" did not mean leaving it, but rather "waste" meant spilling it on the floor. For him, "Oh, gee, I wasted my milk," meant "I spilt it."

These are the kinds of problems that black children encounter on tests. In addition, slang may produce confusions in meaning. For example, if you ask a black child what a fox is, the answer may well be "a pretty girl." However, the multiple choice answers to the question, "What is a fox?" may include "a four-legged animal" but is not likely to include the option "a pretty girl."

There are also phonological differences which cause problems on tests. For example, in the language of many black English-speaking youngsters, when you have final consonant blends, the different sounds usually disappear. Thus, "bowl" meaning the vessel for cereal and "bold" meaning the valor of a soldier actually sound the same, in the same way that "lip" and "rip" sound the same to a speaker of Japanese. So when you ask a black child on a test, "Are bowl and bold the same or different?" and he says they are the same, it is difficult to know whether they are the same in terms of his dialect, or whether he does not know the same/different distinction. Equally, when the test uses rhyming and you ask the child, "What rhymes with bowl, give it to me in a sentence," and the child says, "Oh, I know what rhymes with bowl, cold rhymes with bowl, like an ice cube is cold."

Again there is a problem, because what usually happens is that the teacher says, "No, cold doesn't rhyme with bowl, it's the letters." But in fact, rhyming has nothing to do with "letters," but by the time the teacher explains the whole letters thing, the child ends up not only ignorant of what the teacher was talking about, but also confused as to what a rhyme is. Of course, he knew what a rhyme was when he said, "Yes, coal rhymes with bol'."

These brief examples demonstrate that there are cultural differences and they present a major problem for the child when he is confronted

with the tests. All of these examples reveal the important fact that minority children do have a viable culture, though different from the mainstream, and that in most instances standardized tests do not assess that culture and, in fact, measure something very different. To many teachers and minority group people the cultural difference argument constitutes a reasonable justification as to why tests should not be used. Teachers don't like the blame for the child's failure, minority group people don't like the blame either, and in addition they may be embarrassed by the notion of the differences which in the past had been associated with genetic inferiority.

The question remains: Is there legitimacy in dismissing tests because they are in fact culturally biased? Is this reasonable? Perhaps the answer is no. Why do I think the answer might be no? First, let us look at what the test is and what it does. It is true that the test is developed out of the mainstream culture, and that it reflects that culture. It is also true that the test is usually a good predictor of school performance. If a child can master the test his chances for making it in terms of jobs and opportunities are greater. Given that, then what does it mean for a child outside of the culture who takes the test? For him, taking the test is not the same thing as taking a test for a child who is a product of the culture.

A child who is taking the test that is a product of the mainstream culture is really demonstrating how he has learned the culture. A child who is not a product of the culture is demonstrating how he has acculturated: how much he has in fact learned about a different culture, a culture that is usually very difficult for him to learn about because it is frequently presented in a way that says, "The way you are is no good and adopting this behavior is the only way that you can be good."

The tests then assess how well a culturally different child has learned mainstream norms. Therefore, to say that the test is culturally biased and should be thrown out is to cheat the child of the opportunity and the obligation, of the school system to prepare him for the mainstream.

Since the goals of public education appear among other things, universally to call for the preparation of all children to function economically, and socially in society, until those goals are rejected it would appear unreasonable to reject the tests that serve as a barometer as to whether the goals are being met. It is quite like shooting the messenger because he brings bad news. If one of the goals of public education is knowledge of the mainstream culture, then tests that are culturally laden with such information would appear to be appropriate tools. Of course, when using such mainstream ladened tests we must remember when we interpret the results that those tests are indicators of acculturation and not intelligence.

The problem that the District of Columbia Public Schools encountered following the *Brown* v. *Board of Education* decision is an apt

analogy for the "barometer" approach to tests. The District of Columbia had a segregated school system in 1954. When the *Brown* decision was announced, the school board immediately called a press conference, and issued a policy statement which among other things said the segregated school system would be dismantled and instead D.C. would operate a "color blind" school system. D.C. officials claimed they would no longer be interested in, or keep track of, who was black or who was white. However, the school system quickly rescinded that policy statement when they realized that in order to assure compliance with the *Brown* decision, they had to know who was white and who was black. Without reporting such information there would be no way to document whether or not students and personnel were in integrated settings. In the same vein, tests results assure that the minority child does not become the "invisible man." They assure an occasion for the public to demand that the school make new efforts to effectively educate children. They assure that the problems of such children will not be swept under the proverbial "rug," and their educational needs ignored.

Some people have gone a step beyond just throwing out the tests and have advocated that one should create alternative tests that focus on the child and his culture. Unfortunately, there have not been many attempts at this, or at least there have not been serious attempts at this. One such test, the "Black Intelligence Test for Cultural Homogeneity" (BITCH test), developed by Robert Williams is a multiple choice test which is based on slang or on culture specific items, such as, the time necessary to prepare chittlins, or the names of prominent black performers. While these kinds of questions are clever, and may certainly demonstrate that if you don't know the culture you can't pass the test, they do not appear very valuable for determining if the school system has reached the goals that it set for itself. Until the goals of the schools are changed, it seems to me that it is legitimate to ask if those goals are being met.

Where does all this leave us concerning tests? Ten years ago we had tests, and we had large numbers of minority group children failing those tests. Today, we still have the tests, and the children are still failing. In addition, the controversy over what to do with the tests and with the educational failure is still with us, consuming a considerable amount of time and energy of citizens and educators.

The cultural bias controversy of tests has led spokesmen to suggest that different tests be used or in fact, that testing be abolished. I have attempted to present an argument against those alternatives because (1) the school is obligated to prepare children to function in the mainstream, and (2) test results provide a vehicle for interested citizens to rally around, in an effort to keep the educational system accountable.
18 While test results may not lead to accountability or educational inno-

vation and change, they do provide one instance to assure that minority group students and their interests are not ignored.

This paper has dealt with the legitimacy of cultural differences and why they must be recognized and understood so that tests are not used incorrectly. The cultural differences that appear in the test situation are magnified many times over in the context of the educational system for minority group children. We must recognize the pluralistic society in which we wish to have our students function and use that knowledge not only to intelligently interpret test data, but also to be effective within the day-to-day educational process. It makes good sense and follows the age old educational maxim: start where the child is. Give credence to his culture, legitimize his language, use it to teach him new behaviors—not because the new behaviors are better, but rather because he will be expanding his experiences and will be provided with additional knowledge necessary to function in those different environments. Such an educational approach to cultural difference is a pluralistic one. It does not demand assimilation. It does not say to the child "behave like me and you are good, abandon your cultural and linguistic behavior, it is bad." Instead what it says to the child is that in certain situations standard English is desirable, in other situations black English is desirable. If he learns mainstream norms he can decide if they are valuable to him, and if he wishes to function in the mainstream.

At this point most educational experiences presented to children do not give them that choice. The educational system usually tells children, directly or subtlely learn my way (and, thereby, prove you are human) or forget it. Many children are alienated by such a degrading process and are more and more responding to the system by saying, "forget y_u"—an unhappy solution for all.

VALUES AND PURPOSES IN THE USES OF TESTS

SAMUEL MESSICK

A distinguished Princeton sociologist once remarked that whenever there are several parties engaged in an enterprise and that enterprise subsequently fails, the finger of blame always points to the weakest party. In education, as an enterprise that is often considered to have failed or be failing, the blame has been directed toward different groups at different times. For many years, the problems were considered to be the fault of the student because he didn't work hard enough or he wasn't trying. Then for awhile, the finger of blame pointed to the parents for not providing a good home life or the types of resources that would benefit the child in his schooling. At other times, the schools were blamed for not doing their job, and under this attack the teachers and school administrators turned to social science for assistance in improving their performance. As a result, the finger of blame now points towards social science.

The important thing to recognize is that what is really being discussed in these arguments about blame is not the issue of culpability but the issue of weakness. These types of arguments, which are based on assumptions of strength, weakness and power, are political rather than scientific and substantive. Inherent in political arguments, at least tacitly are always ideological issues. My argument is that this is appropriate and that the ideological problems are the ones we should be discussing in the issues of education and testing.

Much of the current controversy surrounding testing is about tests as instruments, instrumentalities, methods for doing certain things or achieving certain goals. The arguments tend not to be about the goals *per se;* they are about the methods. Today I would like to put forth a general principle that underlying every methodological controversy is a substantive or philosophical issue that is not resolved (or perhaps resolvable). If we expect to move forward, we cannot do so until these questions are addressed.

Whenever a test is being considered for a particular purpose there are always two questions that must be asked. First, is the test an adequate measure of what it is supposed to assess? Second, should the test be used for that purpose? The first question is a technical and scientific one which can be answered by examining the evidence about the psy-

chometric adequacy of the test and more important, by examining its meaning. We have responsible standards for making that type of examination. We know how to gather scientific and technical evidence to determine if tests measure what they intend to measure. Whether or not we gather that information in each instance is another question, but we do, in fact, know how to do it.

The second question, should the test be used for the proposed purpose, is an ethical question, and must be answered by examining and appraising the potential social consequences of the proposed use. For example, we must examine the consequences that are not only directly entailed by the use of the test, but also those consequences that are made more likely by the use of the tests; the by-products, side effects, and delayed effects that might be operating from its use.

We should not delude ourselves, as we frequently do, that the answers to the first question, is the test any good, are adequate answers to the second question, should it be used. Only in the negative case, where the test is so bad psychometrically that it should not be used for anything is the second question answered along with the first. But most often, the two questions are separable and should be separated because the nature of the evidence and the argument is quite different.

I would like to make a few remarks about the first question, that is the meaning and adequacy of the tests, because the issue of cultural bias is one of the important issues involved. But most of my remarks will deal with the second question because I think that the fundamental questions are really ideological and value based.

With respect to the issues of test bias and test fairness, cultural fairness and cultural bias, the key question is what is the meaning of the test? What does it really measure? One of the confusions in this field is a confusion between competence and performance. Much of the evidence that we use to support the interpretation of a test is really evidence in support of interpreting correct answers or high scores on the tests.

If you pose a task to a student and the student can perform the task correctly, it is a reasonable inference to say that he has competencies, skills and processes required to do the task. The problem arises in interpreting low scores or failure in performance. Unfortunately, the simple interpretation is that if you do not perform the tasks correctly you are incompetent, and that you lack the skills necessary to perform the task.

My point here is that it is very important to discount many other plausible interpretations for the lack of performance and not simply assume the results demonstrate incompetence. For example, suppose a student trying to respond to a dictated spelling test gets all the answers wrong. One possibility is indeed that he does not have the competencies required in spelling. But there are many other possibilities: he

may not have been paying attention, he may not have been motivated, he may not have heard well, or he may be deaf. Other possibilities such as these should be taken seriously and discounted before simply interpretations of the failure are made.

More fundamental to this kind of question from the point of view of cultural bias, is the question of what is the nature of intellectual competencies themselves. Where do the competencies originate? Are they innate or genetically determined? Or are they a product of learning and experience, including educational experience, and therefore products of culture?

Most of the modern theories of intellect are fundamentally interactive, that is they postulate that children are born with certain simple, primitive processes. Piaget identifies two such simple processes: assimilation and accommodation. Cattell emphasizes one major basic process, a kind of primitive reasoning ability, which he labels "fluid intelligence." Guilford talks about five such processes, including cognitive awareness, memory skill, convergent and divergent production, and evaluation. None of these basic processes, however, constitute what is meant by intelligence.

The intellectual functioning of the child develops by the application of those very primitive, sometimes sensory-motor, processes to the structured environment in which we live. The intellect or intellectual functioning is neither the information that is in the environment nor the basic processes that the child is born with, but the genuine interaction between the two. From this theoretical perspective, arguing as to whether intellect is determined more by innate process or more by environment and experience is a little like arguing as to which is more important in determining the area of a rectangle, the length or the width. Area cannot be determined without both measures, just as intellectual functioning, according to these interactive theories, requires both processes.

Thus, in a very fundamental way, intellectual functioning is a product of experience in interaction with basic and primitive processing mechanisms, with more complex processing capabilities growing out of the interaction. To measure or assess those competencies which are a product of experience in a culture, the test that measures them directly and appropriately will likely be laden with culture. For many of the competencies which we assess, then, the tests we use are culturally laden, which is somewhat different from saying that they are culturally biased.

From this point of view, the search for so-called culture fair tests of intelligence and achievement may be a snare and a delusion. What does it mean to measure a competency that reveals no cultural or experiential difference? The kinds of competencies assessed would likely be directly related to basic and primitive processes like those present

at birth. These are the competencies which are not subject, in any simple way, to cultural or experiential influence. Competencies and skills that are not greatly affected by cultural and educational influence, generally tend to be more in the sensory and perceptual realm than in the higher levels of cognitive functioning. So it is difficult to argue that tests should show no cultural differences and simultaneously argue that the fundamental characteristics that you are really concerned with are subject to educational and environmental influence.

Another aspect of this problem is the distinction I mentioned between competence and performance. Dr. Baratz indicated by several of her examples that you can have situations that are essentially test items or tasks that the child does not perform well on, and she urged that instead of inferring that the child has no competence, test users should seek other situations that might reveal the competence. Rather than being an indication of competence, the test score might be an expression of the interaction between task and context. And so, although the competence may not be revealed in the test item that is used, there may be other situations which would clearly reveal the competency.

Much of the current work on the cultural context of learning and thinking has made this point dramatically clear. However, we should keep in mind that there are different reasons for testing. Sometimes we are interested in assessing a competence in order to direct appropriate educational experiences that would foster its growth. In this case rather than emphasize different concepts or different competencies for different cultural groups, it would be more appropriate to focus on the same competencies by using different materials, in different contexts. But at other times the society is concerned with level of performance on specific tasks in a specific context. In this case, the issue that must be confronted is that it does not matter whether the competence for that task can be revealed in some other situation, if performance cannot be demonstrated in the socially valued context in question. This is part of the controversy because it is necessary to define clearly the tasks and settings for which performance must be demonstrated and the range of alternative contexts in which demonstrated competency is also an acceptable indicator of proficiency.

For example, for a French speaking child, it does not make much sense to present a task that requires the use of English, if the child cannot speak or understand English. The child would likely fail such a task. In such a case, it is inappropriate to describe the child as being linguistically incompetent because clearly the child is competent in French. But if speaking and understanding English is an important and required task, the child's linguistic competence in French is insufficient. Here the critical issue is the inability to perform in English. It is of course important for test users not to confuse the child's inability to speak English with a general inability to use language; it is also impor-

tant for educators to build upon competency in French to achieve proficiency in English.

Let me now turn to the second question of test use, because it is the more interesting and more intractable one. This issue essentially requires that in order to make a decision about the appropriate use of a test one must inquire into the social consequences of the test and evaluate those social consequences in terms of social values. The complicating feature is that value issues differ for the different uses of tests. There are several broad areas of test use that we all know about: We use tests for selection into particular programs or institutions, we use tests for guidance and placement, we use tests for learning and instruction, and we use tests for evaluation—for evaluating programs in order to improve them, to certify that they are acceptable, and for general purposes of accountability.

Some of the difficulty arises from the fact that several of these areas of use, particularly selection and accountability, are adversarial in nature, and whenever you are in an adversarial situation, you are bound to have argument and controversy. The adversarial quality, however, is not intrinsic to the use, and it is possible to develop accountability systems that are not adversarial. People are very suspicious of them, to be sure, but it is possible.

One problem in the use of tests in support of any of these particular functions is in determining what is fair and unfair, or just and unjust. Another difficulty that arises is that the judgment of fairness is clearly a function of social values and societal preferences. There are many broad approaches to educational systems that exist in the world and the way we would use tests in each of these systems would differ fairly dramatically. For example, we might have an educational system which is geared toward maximizing excellence and maintaining high standards for adult functioning. This is the kind of system that by and large exists in the United Kingdom and some other countries in Western Europe. Or we might have an educational system that is oriented toward sorting students in ways that would maximize the distribution of talent to meet societal needs. This is the kind of educational system that by and large exists in the Soviet Union and some Scandanavian countries. Or we might have an educational system that is geared toward maximizing individual choice, and this is an educational system which, at least in rhetoric, is favored by the United States.

In the first case of maximizing excellence and maintaining standards, we might have an educational system that selects and treats students on the basis of their high ability or high achievement in areas of their individual interest. In the second educational system, the distributive one for societal needs, we might select and treat students who have high specific ability required by socially important fields regardless of personal interest. In the third system, one that favors individual

choice, we might develop procedures for providing students and institutions with information that could affect a realistic and satisfying match between ability, interest, and performance requirements.

The way we would use tests in those three systems would be quite different. Many different educational systems, and hence the testing procedures that are meant to serve those educational systems, could be set up depending on the particular principle of fairness or justice that a particular ideology would favor. And there have been several such principles that have been enumerated. Each of these principles of fairness, in fact, can be associated with some famous philosophers in the history of Western philosophy. One of the principles that we often hear discussed is the principle of equality.

Consider an educational system where equality is the main principle for distributing scarce resources and for making available positions within particular programs of treatment. That seems like a fair principle. However, administering a principle of equality becomes complicated when you start asking questions about equal in terms of what: Equality of treatment? Equality of opportunity? Equality of outcome? At a superficial level, for example, universal, compulsory education is an example of equality. But, should a teacher's time and attention be distributed equally among students?

Another principle for distributing goods, not in the economic sense of materials but in the ethical sense of valued things, might be in terms of ability. Maybe a teacher should spend more time with able students because they can benefit more from the attention and soar in the development of their particular abilities. Effort is another principle for discussing equitable distribution. Maybe a teacher really ought to pay attention to those students who are motivated and work hard, since that is where the payoff is most likely to be. But, in both of these cases one can argue the adverse: That more time should be spent with the students who are low in ability, or who are not motivated, in order to change that situation.

The combination of ability and effort is sometimes talked about as merit. But there are other principles. In addition to principles of equality, ability, or effort, fairness may be defined according to a principle of need: We should distribute resources, attention, time, and effort, to those students who need it, more than to those students who do not. Another principle of fair distribution would be interest and desire: give resources to the students who want them—not everybody wants everything. Another principle might be luck, or random chance. Since with scarce resources you cannot be fully responsive to everybody, use the luck of the draw. Such random distributions based on a lottery have been seriously presented as procedures for selection in higher education. Rather than use tests to select students in terms of ability, in terms of need, or whatever, why not just pick them at random?

Another principle is one of status. You provide educational programs and resources to students in terms of their status: defining status in whatever way is appropriate. We have educational systems in this country, for example, based upon status determined by wealth. Students who have the financial resources can choose to go to private schools or even have private tutors. A final principle, of course, is that of utility. This would imply an educational system set up, in some sense, to do the greatest good for the greatest number.

There are two points I would like to make about these principles. First, you can develop various mixed models by combining the principles, as we have already done in many of our educational institutions. For example, combinations could be used that require certain minimal standards for entry into a program or for the availability of certain educational treatments, but above that level principles other than that of a social minimum would be used. You might even use a random principle, such that entry into college or into particular kinds of programs might be based on the establishment of minimal standards below which a student would have difficulty performing adequately; but above that minimum, rather than selecting the very best, selection would be random. If you take that seriously, as many people have, you very quickly confront the notion that maybe the minimal standard has to be different for different fields. It may be higher for fields in which the public is served. You may even want to change models as a function of the risks involved. Thus, in a field where the risks of failure are inconsequential or perhaps only monetary, low minimal entry standards might be used. But in fields such as surgery or air traffic control where the consequences of failure are great in human terms, you may want to use an ability or ability/effort model.

A final point is that each of the principles mentioned is fair under some specific circumstances, and arguments about what is fair and unfair frequently are arguments about what conditions are or ought to be. This suggests that instead of always arguing about what is fair or unfair, we ought to be spending some time and attention considering the nature of the conditions, particularly those aspects of the conditions that are impossible or very difficult to change, in contrast to those conditions that are possible or easy to change. Since conditions are rarely immutable, we could then consider the problem of designing conditions for which particular principles of fairness would be ideal.

The use of tests to support any educational system or to support any of the principles of fairness or distributive justice that I have just outlined is usually justified in terms of improving social efficiency. At least that is the justification on the surface. What I have tried to make clear is that very frequently the uses of tests also represent choices with respect to social philosophy that must ultimately be justified in terms of social values.

THE ROLE OF TESTS IN POLICYMAKING

JEROME T. MURPHY

My purpose today is to discuss assessment and accountability programs at the state level. During the last six or seven years, there have been a number of attempts to use objective test score data as a basis for making state level policy decisions. At least 23 states have passed accountability legislation of one kind or another, and several other states have become involved in assessment without legislative mandate, so that currently, more than 30 states have ongoing, operational state assessment programs.

Generally, these state programs and their respective legislations have three common features. First, they rely heavily on the use of standardized tests—either norm or criterion reference tests. Second, virtually all the laws discuss "efficiency and effectiveness," despite the fact that these key words have not yet been adequately defined. And third, the legislation has placed a high premium on the "rational process." This process requires first, the establishment of clear goals that are agreeable to everyone involved, then the designation of specific objectives, followed by an exploration of the various alternatives, and then presumably, the selection of those projects and programs that are best designed to improve education. This "rational process" is consistent with how statewide assessment has been approached in New Jersey, as well as with the state's administrative guidelines for the implementation of "thorough and efficient."

During the last few years, I have had the opportunity to visit about half a dozen states which are involved in assessment and accountability, to observe the actual implementation and operation of these various programs. In addition, I have tried to keep abreast of the recent literature which deals not only with what accountability and assessment "ought" to be, but also with what happens when these programs are actually put into operation. From these two sources, I would like to make some tentative observations about the impact of these assessment programs, and second, to briefly discuss some of the possible implications that these observations might have for New Jersey.

My first observation is that, by and large, the data generated by these programs simply has *not* been useful in making policy decisions at the state level. In the case of norm reference tests, the usual product

is a large amount of data that indicates the status of school districts in relation to one another. But the fact that School District A may be doing better than School District B, does not provide any type of explanation as to why this is happening. Does this mean that A is more efficient and more effective than B, and if so, why? This is the type of explanation that legislators need in order to make informed policy decisions and yet the tests themselves do not provide the means for answering that question.

Criterion reference tests, which are the kinds of tests that most states are currently using, have similar drawbacks. One major problem is that the data presented to policymakers is usually so obscure that it is difficult to determine what to do with it. For example, the Early Assessment Program in Florida was designed to give state legislators the kind of information that would be useful in making decisions about the allocation of funds. The first recommendation (and the prime recommendation upon which the members of that body were to act) reads as follows: "Phoneme-grapheme correspondences should be taught without regard to position unless position is important to a specific graphemic representation of a phoneme." Perhaps this kind of data might prove useful to somebody who knows about reading, but its value to Florida legislators who were trying to use it as a guideline for education policy decisions is dubious.

Similarly, in other cases the data generated have not proven to be very useful because the results have been so voluminous. There is no effective means of dealing with the plethora of data that results from these types of testing programs. For example, when Michigan initiated their statewide assessment program, they tested fourth and seventh graders in all schools of the state using norm reference tests. What resulted from these tests was a composite score for both the fourth and seventh grade students for each school district. Such reporting allowed school administrators to compare their school district with others on these two composite scores.

However, when the state, in 1973, replaced the norm reference test with criterion reference tests, they replaced the two composite scores with data on *all* the different objectives that were being tested in the new program. In that year, the state tested for 126 different objectives, in each of its 3000 schools, which in turn produced more than a third of a million achievement facts about how the children were supposedly doing in the education system. There is a Michigan legislator, Senator Bursley, who was responsible for the passage of recent school finance reform legislation, and who used to refer to the earlier composite score reports for each school district as a Sears and Roebuck catalogue. Although I did not talk to him after he received the new data, I can just imagine what he must have gone through while trying to sort out this third of a million achievement facts! Needless to say, with this re-

porting system school administrators were no longer able to make interdistrict comparisons.

Saying that test data have not been useful for making policy decisions at the state level is not to say that such data have not been used. For example, I have read material put out by the New Jersey Education Association that prominently displays the rather ingenious uses of the data by real estate agents. The Michigan Chamber of Commerce has produced several publications which have used the test score data as an argument for *reducing* school funding. The Chamber of Commerce argued that there was no need to spend more money on education since the assessment data indicated that districts with high per pupil expenditures did not necessarily produce students who performed best. In short, the data have been misused.

My second general observation, which also grows out of my experience in Michigan, is that establishing minimum performance objectives is one of those ideas that sounds great until you try it. Rather than providing objective data for decision-making; Michigan's first year program produced an enormous amount of confusion about what the right minimums ought to be, at what level they ought to be set, who should set them, how they ought to be used, how extensively they should be measured, and how they should affect classroom life. If the minimums were designed to improve the education of children in Michigan, I have discovered no evidence that demonstrated a direct impact.

Third, while it is true that the accountability programs have not generated a lot of light, it certainly is true that they have generated a lot of heat. In the case of Michigan, everybody from the National Association for the Advancement of Colored People, to the Civil Liberties Union, to the teachers' organization, to the administrators got involved in the problem. Enormous battles raged primarily between the teachers' association and the Chief State School Officer, John Porter, who was the major proponent of the accountability program. A newsletter that was put out by the Michigan Teachers Association called "Accountability Alert" is one indication of the acrimony of the debate. Widely distributed across the state, it contains a cartoon with a picture of John Porter on a steamroller with kids lying in his path. The caption reads: "Steamrolling along with John Porter's education accountability model. Guaranteed to iron out all problems." If anything can be said about this debate, or at least the approach to the debate by the teachers' association, it is that is has all the subtlety of a sledgehammer.

Having made these general observations and having discussed some of the problems with the assessment and accountability programs in the states that I have studied, let me now discuss what I see as some of the reasons underlying these particular results. One of the problems is

clearly technical. In the case of the norm reference tests, the kind of data that are generated are not the kind of data that indicate qualitative differences between schools or school districts. Although the data can tell how well students are reading in different school districts, it is impossible to determine which part of the test score is due to either good or bad teaching, or which part is dependent upon other factors like intelligence, or socio-economic background. This type of technical problem reminds me of the testimony a principal gave to the Michigan Board of Education. She explained that at one point in her career she had been the principal of two different schools. In an attempt to equalize the quality of the schools, she had initiated new ideas and had transferred her best teachers into the lower achieving schools. But despite her efforts, when the test scores were released at the end of the year, the "worst school" continued to have the lowest scores. Obviously, although the tests were reflecting some kind of a problem, they did not aid in determining its origin.

In the case of objective reference tests, some technical problems exist because this is a new area which is still in the process of being developed. The state of the art is still rather primitive, and while technical changes will be made in the next few years, in my judgment, there are currently too many problems that have to be worked out before objective tests can be useful for policymaking.

Another reason for some of the problems I have identified is conceptual. A basic assumption behind the programs, is that it is possible to get agreement between the people who are involved, concerning the establishment of goals and objectives. While such consensus is admirable in principle (and indeed it is possible to get people to agree on goals, if only at a very global level), when you attempt to describe the specific details of the objectives these goals exemplify, the experiences in Michigan and other places suggest that controversy results.

One of the reasons for this is that people naturally have different conceptions of what schools ought to be doing. Some see the school's main function as bringing up the level of reading test scores, others see it as developing a sense of self worth in children. In addition, I am not sure that people really know what their goals and objectives are in a complicated social and moral area like education. Consider the following quote from a scholar in the field:

> *Not only do our social ends and values conflict, but being quite subtle and complex they are exceedingly difficult to specify. We simply cannot determine in the abstract our ends or values and the intensity with which we hold them. We discover our objectives and the intensity that we assign them only in the process of considering particular programs and policies.*

The third reason for some of the problems in the assessment and accountability programs concerns the knowledge base in education. Let me illustrate my point by making a comparison between a state's education office and its highway department. If we were talking about establishing standards for the highway department, it would be a relatively easy matter for a group to decide on what the minimum objectives ought to be. For example, one minimum objective might be that the highway should last for at least 20 years. Another objective might be that the cars don't leave the road when they go around turns. Another objective might be that when it rains water puddles are not created in the middle but instead flow off to the sides. Similarly, if you were to talk about ways of achieving those objectives, it is fairly clear how to proceed. For example, we have the technical knowledge as to the mixture of cement and sand required to have a road last a certain number of years. If you want to control water flow, we know what the pitch of the road ought to be. If you want to prevent people from going off the road, we know from engineering studies how to bank the curves.

This kind of agreement on what the objectives in education "ought" to be does not seem to exist and certainly the kind of knowledge about how to achieve any such objectives does not exist. In short, trying to apply the highway department model to education runs into enormous problems because the knowledge base is simply different.

Finally, the debates concerning both the problems in accountability and assessment programs, and the reasons for the problems, are normally presented in educational and technical explanations thereby excluding many policymakers. But, at bottom, the reasons for the problems surrounding assessment are political. People simply recognize that collecting test score data for policymaking purposes is a patently political activity, particularly when there isn't any scientific base for choosing the objectives for testing. Information can be and will be used as a political tool in the allocation of resources and in the allocation of jobs. Because of the stakes involved, people are going to fight over who is tested, what is tested, when it's tested, by whom, and how the data ought to be released.

With this "positive statement" about the state of the art of testing and about my experience with accountability schemes, let me now turn to what all this means in general, and then discuss some of the implications for New Jersey. First, in regard to testing, it seems to me, that we need to establish more reasonable expectations for the use of testing in the policymaking arena. With a better understanding of the limits of tests it is conceivable that the demagoguery that has marked much of the debate might possibly be eliminated, or at least reduced.

Assuming that this happens, it seems to me the tests can be useful in providing census-like data on the status of students or school districts,

and on trends in educational attainment. However, tests are unlikely to provide useful data on the efficiency and effectiveness of schools. An example of the census-like use of tests can be seen in the experience of the national assessment program. In the early days of the program, national assessment was under attack because nobody knew exactly what use to make of the data. Now that it has been in existence for several years we are in a position to compare results from different years. This comparison provides useful data on how things seem to be changing. The Scholastic Aptitude Test data provide another example. Although we don't know why scores move in a certain direction, the change in the Scholastic Aptitude Test scores can act as a bellwether, alerting people to a potential problem and the need for investigation and improvement.

As to New Jersey, I have had an opportunity to examine some of the materials that have been prepared on the establishment of minimum performance objectives. While I have a great deal of sympathy for what people are trying to accomplish with minimum performance objectives, I nevertheless would be quite reluctant to support the various proposals that I have seen. These include the one put forth by Mr. Tractenberg and the Educational Law Center which calls for the establishment of statewide minimum performance objectives. Similarly, the alternative suggested by the Department of Education, to have minimum performance objectives established by various localities, for many of the same reasons, does not make sense. And, finally, looking over the administrative code, the proposed process of getting every school district to go through a long, horrendous procedure of establishing goals and objectives strikes me as a waste of time and resources unless the goal is one of two things: to provide a symbolic bow to the improvement of schools, or to get people in the short run to feel better about their schools. But in the long run, one must wonder. There is very little evidence that this whole process has anything to do with decisions about the allocation of resources. Those of you who are involved in organizations and have gone through this goal setting process, must wonder how much impact such activities have on what happens. My impression is that while the goal setting goes on in one place, the real decisions about education are being made somewhere else.

One of the problems with doing research is that, although it might be easy to analyze the problems, it is still impossible to figure out what to recommend! Let me suggest some very tentative alternative ways to generate information that might be useful in improving the schools. These are programs that I only know a little about and I am not suggesting that they necessarily be implemented, but it seems to me that they are alternatives that ought to be examined.

32 One program is in Kalamazoo, Michigan. In that case, local officials

have not relied just on test scores to gauge teacher effectiveness, but have tried to collect a variety of other kinds of information as well. They have students rating teachers, administrators rating teachers, parents rating teachers, and teachers rating other teachers. The idea behind the Kalamazoo program is that the opinion data, combined with the test score information, can provide valuable insights into teacher performance.

Another alternative for providing information is through something called "annual school reports." This idea is currently being implemented statewide in Florida and Ohio, and also in several school districts across the country. The concept behind these reports is that each school in the district ought to provide parents with a variety of information in addition to test scores, that can help show what is happening in the school. The principal, the teachers and parent advisory council members are responsible for compiling an annual report that tells parents what is going on, what the problems are, and how the school is dealing with them. It is hoped that this kind of qualitative data, in addition to some objective data, might provide parents with information they can use in making judgments about whether or not their school is working.

A third approach to generating useful information is being followed in the state of Colorado, which has had an educational accountability law since 1971. In that case, however instead of relying on statewide test scores, legislators have authorized local accountability committees composed of teachers, administrators, taxpayers, students, parents, etc., to act as a watchdog over the schools, to collect a variety of information about what kind of progress is being made, to explore how things can be improved, and to feed that information back to the local communities so that some decisions can be made about reforming things.

Having offered those alternatives and sensing their high level of acceptance by this group, let me close by simply repeating my main point: there are alternatives to tests that could provide some useful data in helping policymakers make decisions about the improvement of schools.

TESTING AND ACCOUNTABILITY: A POLITICIAN'S PERSPECTIVE

SENATOR JACK GORDON

I have been trying to determine in what direction this discussion would be moved so that it would have some relevance for legislators and other citizens concerned with education. Consequently, I would like to talk about some of the major problems with public education and schools as they relate to accountability.

As was said earlier, test results do not "speak for themselves." The results and assigning responsibility (blame) for them. When I walked in this morning, Joan Baratz was talking about the question of cultural bias in tests. Cultural bias is supposedly one answer to the question of blame for the failure of children which, of course, is a typical teacher way of talking about children failure: the *children* have failed. Some of us who like to look at the educational process might want to explain the test results of culturally different children as teacher failure. It is this notion, that testing is in some fashion related to teacher performance, that upsets teachers. However, explaining test results is not an easy task. There is probably enough blame in the whole situation to be shared by all the parties. What we ought to look at in thinking about testing is what we are doing in the whole educational system. That system is built, as anyone would expect in this present day American culture, on an essentially industrial model. We have regarded our industries as the most efficient form of production (of course you only have General Motors to validate the fact that General Motors is efficient!). Nonetheless, we do look to an industrial model for our school system.

The industrial model of education works like this:

- We pass children, as units, through a sequential kind of production line, and we pour in certain stated pieces of knowledge, classified in arbitrary ways.
- The diversity and arbitrary nature of the knowledge sets up all varieties of professionalism, because only professionals can comprehend those particular pieces being dropped into a child's head.
- At the end, we come out with a product.
- And finally when we attempt to impose some kind of quality con-

trol, some of the "rejects" get pretty upset about it. In schools, as opposed to industries, the products can talk, and that is the difference we are discussing.

There are, of course, many ways to hold schools accountable. Test scores are one measure. We could validate schools on the basis of how many students score over a certain point on a College Board admissions exam, or how many got into prestigious schools, which is a way, for example, that prep schools regard their record. Any prep school brochure you read emphasizes the colleges to which their students were admitted. And, if there are a sufficient number of students admitted to Harvard or Princeton, then presumably it is a particularly good school. If that is what you are educating for, that may be as good a way as any to evaluate a school.

I would like to suggest that there are alternative accountability notions that don't involve tests at all. There are methods to evaluate schools without tests. For example, a very simple way would be to consider a school's drop-out rate. If students are supposed to be educated for twelve years, in a twelve-year sequence, and a number of students, for whatever reason, reject continuation of school for what they perceive as a better alternative to staying, that says a great deal about the school.

I recognize the economics that are involved, but I don't know whether the New Jersey Education Association would be willing to test the validity of the high school education in New Jersey by lowering the school leaving age to, say twelve, and observing how many students showed up for high school. Or how many of them would put their jobs on the line based on how attractive they could make high school if attendance were voluntary. It is not an unreasonable kind of choice, especially in view of the drop-out rate in the schools.

In Florida, we tried to address this problem in a somewhat different way in our finance formula. Initially our formula weighted the last two years of high school by an additional ten percent so they generated 110 percent of the average full-time equivalent dollar. This was done on the theory that such a system provided a financial incentive to the high school to keep people there longer. High schools became a big accounting problem because they are three year high schools, and kids move back and forth across artificial grade units. But there are ways of playing with finance formulas that enable schools to receive more money if they raise their retention rate (without hiring too many policemen to keep the kids in school). This is one possible way of dealing with an educational problem in non-educational terms.

But people are asking, and legislators and the people who provide the money are being asked—what are we getting for our money? And in one way or another that question must be answered. What happens when you are faced with level or declining revenue, an inability to

raise taxes, and a federal government that is moving further and further away from taking care of human needs in favor of antihuman needs like armaments? Those people concerned about special social problems such as crime, are faced with dilemmas such as—Do you build prisons? Do you build schools? To what extent are you going to support law enforcement? What are you going to do in a country that is really in a depression, or close to a depression, with very high unemployment rates that increase the demand for social services?

Because schools get the biggest chunk of money at the state level, and a very big chunk of money at the local level, all levels of education are being questioned. The amount of available money is limited and the schools have to answer what they are giving people for their money or the decision-makers will spend that money on something else. I think school people have to understand this attitude, and understand that they don't have to take the questioning personally, but nonetheless they have to answer the question.

Dr. Murphy spoke earlier today about programs in Florida. I would like to expand on the socio-political context in which accountability is discussed in Florida. I have been in the state legislature for just three years. This will be my fourth session when we meet in April. I am somewhat familiar with what happened before I came into office. An important piece of information on Florida is that reapportionment caused a dramatic shift in political power in Florida in 1967-68. Where my county, for example, had one state senator and three representatives in the early sixties, it now has eight senators and 23 representatives—a 20 percent share of the legislators in both houses. Consequently, urban areas were concerned about receiving some kind of urban equity across the entire range of state services. As we looked at education we wanted to see some finance plan that would aid urban areas. Florida incidentally has just 67 counties and 67 school districts —one of the two states that has county wide school districts. It has elected school boards, and the constitution permits them to assess up to 10 mills on property. Statute now limits them to 8 mills. But, since the constitutional limit is 10 mills, that is the financial basis of the local effort for schools. Local effort is now about $500 million in a $1.6 billion appropriation.

Counties have some ability to levy more than is required by the state. They have a leeway of about one and a half mills. We asked the counties to have a plan that would coordinate a comprehensive five year plan for education, which they are supposed to implement. The state department has done a pretty fair job, under significant prodding, to actually review those plans and in some places, to push them to do something better.

In dealing with these issues, up-to-date information is essential. Florida initiated the development of an information system at the state

and local levels that permits a school-based budget so the control of the school is at the school site, with the school principal. Since the state has an allocation formula based on how many students are in such categories as early years, late years, vocational and special education programs and so forth, money is easily allocatable per school. At any rate, the school-based budget allows the principal significant control over the resources. We do not permit principals to have tenure as principals, only as teachers—this is much like the higher education model which contends that tenure belongs with instruction.

The next innovation we made was to mandate a parent, student, teacher and community Advisory Committee in each district, although 25 or so of the larger counties in Florida actually have Advisory Committees in every school. Broward County, that is Fort Lauderdale, has one, as does Hillsboro County where Tampa is located.

In Dade County they have developed a hodgepodge within the various districts. The district of Dade County, with a quarter of a million pupils, has six administrative districts. Within these, there is at least one Advisory Committee per district and some of them are organized with the high school and its feeder schools—a fairly reasonable way of dealing with decentralization. Some of those parents and students and teachers and community people have welcomed the opportunity to participate in the school policy planning and help write the annual report of school progress in which the test results, among other things, are published.

Now, you can make all the technical comments you want to about tests and how they are constructed—and how they might be misused. However, if you are going to sit down with a group composed of students, teachers, and concerned people from the community, who are active because their children go to school or because they are leaders in some community action organization, or organizing Title I mothers, or child care, it is necessary for them to work with some kind of data. They will make good or bad use of whatever information you put in front of them. They can see that the school in their neighborhood and the one in a neighborhood like it have remarkably different scores. Theirs might be better; theirs might be worse. They might say let's go over there, or ask the principal why there is such a difference.

If the principal or teachers say we have this or that kind of problem teaching kids to read, one reaction you get from parents, students and the rest of the community is, "Well maybe we can help." "We can raise the number of volunteers, we can fund, or get funded, a program that would bring significant number of senior citizens into the schools as aides."

In Miami we had one of the first grants under the Older Americans Act to get retired English teachers to help English teachers in high school so they could assign more written work and have somebody

else correct it. As it turned out, the most effective utilization of senior citizens was as industrial arts teachers because all the retired mechanics, shop foremen, metalworkers, woodworkers, etc., loved the idea of coming in and helping students learn.

There are a lot of community resources that are available on a volunteer basis. Teachers don't like to consider volunteer resources because they feel volunteers reduce the number of jobs and strike at their professionalism. That reaction really doesn't bother me because I think it is accurate to say that the movement to involve parents, students and other elements of the community in what goes on in the schools is an anti-professional movement, and I think a justifiable one. I am quite willing to consider myself a militant generalist in this area. I think we have over-professionalized schooling and over-specialized it to a point where if anyone without numerous credentials makes a suggestion he is considered to be interfering violently with the education system!

We need to change what we consider the essential ingredients in education. If we are going to have a democratic society with democratic decision-making rather than an elitist process dependent on professionals to make decisions, there has to be some shared way of looking at education; a moving down to the smallest unit and getting the community involved in what happens; and a recognition that a text book somebody wrote in the college of education is not necessarily the ideal educational tool.

People are not going to spend more taxes on schools or raise teachers' or administrators' salaries, or hire more administrators or teachers on the basis of data and assumptions currently made about schools. I can tell you that from my experience, not only in Florida but from looking around the country.

It seems to me from a philosophical point of view that is peculiarly American, one alternative to the current popular disenchantment with public education would be to democratize the schools and their decision-making process and to utilize the kinds of talents that are available outside of the profession as well as in the profession. I do not mean to exclude professionals from decision-making, but I think it is important that decision-making be shared.

Such an approach is, I think, a fairly accurate statement of the philosophical direction of the Florida legislature. I am glad that Dr. Murphy examined what was happening with test data in Florida. He did a great service, and performed an oversight function that ought to have been done by Florida education personnel. They could have (and should have) gone out into 10 districts, I think 100 schools, 10 schools per district, including the large districts in Florida, and found out that all this testing, the phonemic, graphemic consonants and so forth, was being totally ignored.

Some of that feedback from Dr. Murphy and others into the department of education will change the testing program and improve it. But by and large the decisions about what happens to students takes place in the local schools regardless of what state department administrators may think. Consequently, we have a great body of non-knowledge about what is going on, and it seems to me that more community input and participation is a way to improve schools and help make them accountable so people will feel these expenditures for education are worthwhile.

EVALUATION IN PUBLIC EDUCATION

SAMUEL BALL

My subject for this conference concerns evaluation in public education. I would like to begin by discussing evaluation within a public policy context. It seems to me, that in terms of public programs, there is no option as to whether or not evaluations are to be conducted. Evaluation is as natural as breathing. It is only natural that values are placed on things; either you like them or you don't. It is very difficult not to get some kind of feeling about a program. Therefore, the central question is not *whether* we should evaluate, but rather, *how* we should evaluate.

That question could be asked a little more specifically: Should we evaluate systematically, consciously, in a planned, rational way, or should we evaluate haphazardly, fortuitously, irresponsibly, using questionable evidence? Obviously, my answer is that we should be systematic and more precise about evaluation.

That same question can be asked in other kinds of ways, but for now, let me begin with the assertion that evaluation is necessary, and further that the process of evaluation must take into consideration the various meanings of the term. There are two kinds of evaluations that take place: the evaluation of programs and the evaluation of students. Unfortunately, these two types are often confused and it is not unusual to hear public debates in which one person is talking about one kind of evaluation and the other one is talking about another kind.

As far as I'm concerned, discussions concerning the goals of high schools, and whether or not someone should be awarded a high school diploma, are questions which concern the evaluation of students. These same questions may or may not also involve the evaluation of the high school program, which is a separate issue. In contrast to evaluating students, evaluating high school programs does not require testing every student in the school. Instead, on the basis of a random sample of the students involved in the program, an evaluation of its success could be made almost as accurately as if the entire student body had been tested.

Extra care must be taken when the question of evaluation and testing are considered. First, questions focusing upon the purpose of the evaluation must be asked: Is the purpose to find out more about the chil-

dren? For example, is the evaluation diagnostic? Will the data be used to place individual children into individual programs? Or is it to evaluate the program itself? If the latter is the case, then it is not as necessary to have a highly reliable test, because with a large sample of students, the reliability of results is improved. But, if the questions involved deal with a single child, and on the basis of the evaluation a judgment about that child will be made, then the test to be used must be considerably more precise in order for the judgment to be valid and reliable.

My purpose is to talk primarily about evaluation of public programs rather than evaluation of students. My major focus, therefore, will be the gamut of public programs, from things like education programs in prisons, to high school programs, to public television programs or whatever educational programs are available. I would argue that the essential reason for any kind of evaluation is for decision-making and the decision-makers. The evaluation processes have to be developed and planned with the thought that some kind of decision will be made based on at least in part on the evidence provided by the evaluation. It cannot be demonstrated that the evaluation will assist in the decision-making process, then I am not sure that it should be done.

There are five major purposes for program evaluation within the context of decision-making: "Front-end analysis," formative analysis, "end-product analysis," political analysis and the contribution to basic processes. "Front-end analysis" is evaluation that is used for purposes of decision-making before programs are in operation and addresses such policy questions as: Do you need a program? How big is the need? What is the availability of staff? Formative analysis is used in the program development stage of the policy process. Once a need and the capability to develop the need have been established, there are a lot of developmental decisions that have to be made. Whenever money has been set aside for the development of a program, some of that money ought to be used for systematic work in the development of that program. During this phase of program development, an educational evaluator can be of assistance. There is a third kind of purpose which I call the "end-product analysis," or technically, summative evaluation. This occurs when a program is in existence and in operation and it is necessary to know its effects. A fourth evaluation function which is very important is the political purpose. The final type of evaluation is intended to help extend knowledge in general, as opposed to specific knowledge about a single program.

The chart Purposes and General Methods of Program Evaluation, on pages 46-48, is designed to provide an additional focus for the five purposes of program evaluation. Across the top of the chart are categories which I call methods of evaluation. These are the methods used in developing information and in getting evidence including method-

ological techniques such as experimental study, quasi-experimental study, student assessment, expert judgments, case studies, informal observations, and testimony. I have also included a cross tabulation which indicates how the different techniques and methods are used, depending upon the kinds of purposes the program must serve. You will notice, for example, that surveys are heavily used in some types of evaluations but excluded in others. Depending upon the purposes of a program, it may be quite legitimate to use only informal observation and informal kinds of testimony rather than survey methodology.

The first category, "contributing to decisions about program installation," is "front-end analysis" evaluation. This concerns questions of need, which I have further divided into two subareas of frequency and intensity. For illustrative purposes, suppose that the children in sixth grade in a given school system are reading a little below average. This is a fairly common problem, but is not an intense problem because they can all read reasonably well, even though they are not reading quite as well as the children in another district. For the school district in question, the upgrading of reading skills might constitute a frequent, but not an intense problem. Another possibility, of course, might be that within a school district there are a few children who are deaf and require special services. There are not many deaf children but the need there is obviously intense. In developing the need for given programs some sort of method must be established for dealing with the problem of equating the frequency and the intensity of needs.

With respect to program conception, "front-end analysis" must deal with issues relating to appropriateness and quality. These considerations occur at the very early stages of program planning when a need has been established but there is still some question as to whether a program will be instituted. It is important, at this stage, to make the program goals very explicit. Likewise, it is necessary to establish detailed procedural steps, in order to assure that a continuous evaluation will occur as the program goals are gradually implemented. Far too often programs are developed without adequate thought of how to determine whether the program is operating as planned once it has been initiated. Note too, that program conception and cost, in relation to alternative strategies and feasibility are important. For example, we have all seen programs that have been developed, but are impossible to implement because there are not enough teachers who know the particular subject area, or the cost of the materials is too much for the local school districts.

Consider now Purpose II which is the summative evaluation notion of "contributing to decisions about program continuation, expansion, and/or accreditation." I use the word expansion there purposely. Those of you who remember the early days of Head Start know that it was an example of a program that was started in an expanded form

before sufficient evaluative work had been conducted as to its feasibility and its likely impact.

Where political pressures are not too intense, it is far more sensible and rational to start a program on a small scale, assess its effects, develop it so that it is working well, and then start expanding it. There are, of course, great pressures not to do that, especially at the political level. But certainly at the technical and educational levels, the pressures always ought to be in the direction of careful developmental evaluation prior to massive program implementation.

In industry, there are examples of both approaches and especially of products that get mass produced before they should have been. I suggest that in education the same problem exists. I can remember going into New York City and being asked to have a look at Head Start during its first summer. I went to the headquarters of Head Start and asked where the programs were. The staff responded, "Well, we do have a copy of the locations somewhere but we can't find it, and we can tell you where some of the programs are. Maybe you could go there and have a look." I did, and it was nice, the programs were operating fairly well. But that was a matter of good luck and it could have turned out to be much more disastrous than that.

Another element in formative evaluation (Part II of the chart) is the question of continuing need. The continuing nature of the need itself can be a good outcome measure. With some programs, if the need starts to decline, it might be an indication to a decision-maker that the program is indeed successful. But programs have an uncanny ability to be self-perpetuating, especially if it is providing employment for people. So sometimes it is rather difficult to do away with programs even after the needs have changed. I think that this can be seen in schools all around us—programs that have outlived their need.

Section B of Part II, also draws attention to the global effectiveness in meeting the need, and it includes two aspects of a program which are important—short-term and long-term effects. Most often, when evaluations are done they are short-term. There is rarely money set aside for example, for a five year evaluation study. Typically, the decision-maker wants to know how that program is operating, and wants to know as soon as possible, and the only kind of information that can be provided of the "as soon as possible" type is short-term effectiveness. Yet it is the long-term effectiveness that is more important. Although this is often a fiscal problem where legislatures can only vote money on a yearly basis, this kind of problem has to be addressed and the solution is not yet evident.

Note in developmental evaluation, Purpose II, attention is also drawn to the negative and positive side effects. Far too often, especially in education, there is a tendency to look for only the intended outcomes. Maybe, there is only enough money to look for this aspect but

it is the unintended outcomes that may in the long run be more important. In medical research it is obvious, especially since thalidomide, that the unintended outcome can be by far the most potent.

There are probably children who are equally harmed educationally, but they are not usually seen as "educationally deformed," because there are results of some programs whose side effects are less dramatically obvious. I am quite sure I could develop a program that could get most preschoolers reading, but I am not sure what the cost would be, in terms of negative side effects. It might not be at any cost at all, but it may be, and I certainly think that it is important, in a program of this type, to assess such possible side effects.

For example, with Sesame Street, the purpose was not to encourage children to watch more television. Yet this undoubtedly is one of the negative side effects that occurred. Certainly decision-makers need to be aware of these types of problems in order for their conclusions to be well-founded on facts.

Part of formative evaluation includes assessing the demand and support—popular, political, financial, and professional—for the program. Initially, I did not feel that support was a factor in developmental evaluation. However, I was thinking only in technical terms and in practice, such evaluation is important. In many instances, if the program has support, it is assumed that it must be successful and vice versa.

There are times when it is necessary to find out what the real impact of a program is on the students. This is a situation where some form of testing becomes an important counterbalance to other methods and provides additional pieces of information that are valuable in assessing impact. But, let me stress, that to determine whether a program is working or not, it is not necessary to test every student, or every question, and that a sample of students and a sample of test items will be sufficient.

The third purpose of evaluation listed in the chart is to "contribute to decisions about program modification." This is a form of formative, developmental evaluation. My purpose is not to discuss all these aspects of formative evaluation, because which elements are actually carried out depends on what kind of program is being developed and what decisions have to be made. Naturally, it is important to consider things like program objectives, curriculum content, instructional methodology and the program context. If these elements are not planned and developed, then perfectly good ideas will never get implemented. For example, some of the early studies that showed that when the new curriculums were developed in math, science and biology, much of their alleged success or failure depended upon popular acceptance, and especially professional acceptance. Very frequently, the program was wrongly judged not because of the program itself but

because of the interaction of the profession with the program. The teachers, for example, may not have been teaching the new program the way it was supposed to have been taught, or they simply translated it into the old ways that they had learned to teach. And so the notion of what is actually happening in terms of the processes that are occurring in the early stages of program implementation, are very important to assess.

"Obtaining evidence favoring the program to rally support or to rally opposition" is another aspect of evaluation. Although this is not a scientific notion of program evaluation, it is a realistic one. Some people have thought of this function as the "adversary-advocacy" model of evaluation. The evidence that is needed is collected, but often instead of resulting in a technical report that the evaluator thinks approximates the truth, reports are written that represent a strong brief in favor or against the program, written of course by different evaluators. This leaves the problem to the poor decision-maker to choose which set of evidence he or she feels is most likely to be acceptable. This is the "evaluation" model used in the courts of our land, it is not the model that is normally used in scientific investigation, and many people who argue that this type of evaluation is not within the scientific tradition.

A final topic to be addressed is the question of evidence. I noticed in terms of the program today that the word "tests" keeps coming up. I suppose that this is natural in a place like ETS, but I want to stress the point that test scores are only pieces of evidence that might be used in judging the quality of a program, in deciding whether to expand a program, or in deciding whether a particular program component is working or not. But this is only one kind of evidence and in many of the activities related to the purposes of evaluation that are seen in the chart, there is certainly a role for judgment, especially judgment of experts, and for surveys that do not use tests.

PURPOSES AND GENERAL METHODS OF PROGRAM EVALUATION

■ = likely investigation method

I. To contribute to decisions about program installation

A. Need
- 1. Frequency
  - a. Student
  - b. Society
  - c. Other (e.g., industrial, professional, governmental)
- 2. Intensity
  - a. Student
  - b. Society
  - c. Other

B. Program conception
- 1. Appropriateness
- 2. Quality
- 3. Priority in the face of competing needs

C. Estimated cost
- 1. Absolute cost
- 2. Cost in relation to alternative strategies oriented toward same need

D. Operational feasibility
- 1. Staff
- 2. Materials
- 3. Facilities
- 4. Schedule

E. Projection of demand and support
- 1. Popular
- 2. Political/financial
- 3. Professional

II. To contribute to decisions about program continuation, expansions, and/or "accreditation"

A. Continuing need
- 1. Frequency
  - a. Student
  - b. Society
  - c. Other
- 2. Intensity
  - a. Student
  - b. Society
  - c. Other

B. Global effectiveness in meeting need
- 1. Short-term
- 2. Long-term

C. Minimal negative side-effects

D. Important positive side-effects

PURPOSES AND GENERAL METHODS OF PROGRAM EVALUATION (cont.)

E. Cost
1. Absolute cost
2. Cost in relation to alternative strategies oriented toward same need
3. Cost in relation to benefits

F. Demand and support
1. Popular
2. Political/financial
3. Professional

III. To contribute to decisions about program modification

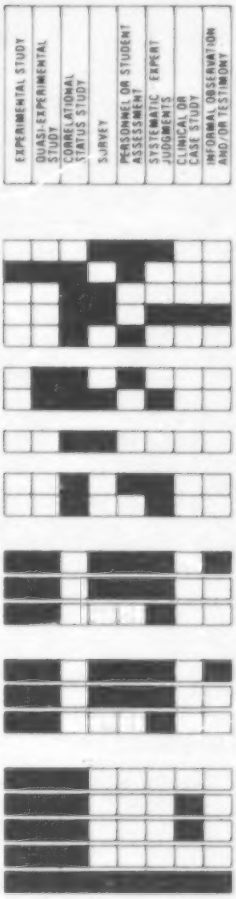
A. Program objectives
1. Validity and utility (in meeting needs)
2. Popular acceptance
3. Professional acceptance
4. Student acceptance
5. Instructor acceptance

B. Curriculum content
1. Relevance to program objectives
2. Coverage of objectives
3. Technical accuracy
4. Degree of structure
5. Relevance to backgrounds of students
6. Effectiveness of components
7. Sequencing of components
8. Difficulty
9. Popular acceptance
10. Professional acceptance
11. Student acceptance
12. Instructor acceptance

C. Instructional methodology
1. Degree of student autonomy
2. Effectiveness of presentation methods
3. Pacing and length
4. Reinforcement system
5. Student acceptance
6. Instructor acceptance

D. Program context
1. Administrative structure, auspices
2. Program administration procedures
3. Staff roles and relationships
4. Public relations efforts
5. Physical facilities and plant
6. Fiscal sources and stability
7. Fiscal administration procedures

PURPOSES AND GENERAL METHODS OF PROGRAM EVALUATION (cont.)



E. Personnel policies and practices
   1. Students
      a. Recruitment
      b. Selection and placement
      c. Evaluation
      d. Discipline
      e. Retention
   2. Instructors
      a. Selection and placement
      b. In-service training
      c. Evaluation for promotion, guidance, retention, etc.
   3. Administrators
      a. Selection
      b. Evaluation for promotion, retention, etc.

IV. To obtain evidence favoring program to rally support
   A. Popular
   B. Political/financial
   C. Professional

V. To obtain evidence against program to rally opposition
   A. Popular
   B. Political/financial
   C. Professional

VI. To contribute to the understanding of basic processes
   A. Educational
   B. Psychological
   C. Social
   D. Economic
   E. Evaluation (Methodology)

NOTE: From *Professional Issues in the Evaluation of Education/Training Programs* by S. Ball and S. B. Anderson, 1975, Educational Testing Service. Project sponsored by the Personnel and Training Research Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-72-C-0433, Contract Authority Identification No. NR 154-357. Reprinted by permission.